# PLATO: Portable Language-Independent Adaptive Translation from OCR

## Quarterly R&D Status Report No. 2

| | |
|---|---|
| **Contractor:** | **BBN Technologies**<br>10 Moulton Street, Cambridge, MA 02138 |
| **Principal Investigator:** | Mr. Prem Natarajan<br>Tel: 617-873-5472<br>Fax: 617-873-2473<br>Email: pnataraj@bbn.com |
| **Reporting Period:** | 1 January 2008 – 31 March 2008 |

This material is based upon work supported by the Defense Advanced Research Projects Agency
DARPA/IPTO
PLATO: Portable Language-Independent Adaptive Translation from OCR
MADCAT Program
ARPA Order No: X 103
Program Code: 7M30
Issues by DARPA/CMO under Contract #HR001-08-C-0004

# EXECUTIVE SUMMARY

This is the second R&D quarterly progress report (QPR) of the BBN-led team under DARPA's MADCAT program. Due to the recent changes in our reporting schedule, there is an overlap in the accomplishments in this quarterly report with the previous quarterly report.

## 1.1 Pre-Processing and Image Enhancement [BBN, Polar Rain, UMD, SUNY]

The goal for the pre-processing and image enhancement task is to eliminate noise artifacts from documents. In this reporting period, we performed preliminary experiments to assess the usefulness of shape-DNA enhancement on machine-print and handwritten images. The shape-DNA approach uses a database of low- and high-resolution shapes and a probabilistic shape-mapping model. The database and mapping are both automatically learned from training data to estimate high-resolution details from low-resolution shapes.

For our initial pre-processing experiments we created a corpus of images from two different sources: (1) 300 real-world documents acquired primarily in Afghanistan, and (2) a controlled document set of 33 images generated by three different Iraqi writers each writing the same 11 distinct documents. Using this corpus of images, we performed two sets of preliminary experiments to separately assess the restoration and cleaning attributes of the shape-DNA approach. In the first set of experiments we assessed the usefulness of the restoration. We first estimated shape databases using 15 of the 33 controlled handwritten documents. Then, we synthetically degraded the remaining 18 controlled documents by introducing breaks in character glyphs and degradations that mimic those observed in documents from the field. Finally, we applied the shape-DNA technique to restore the broken characters. In the second set of experiments, our goal was to evaluate the cleaning of field documents *without* restoring the text. For this set of experiments we did not train on the controlled handwritten data. Instead we used shape databases trained on machine-printed Arabic documents.

We assessed the effectiveness of the restoration and cleaning processes through a visual comparison of the output with the input image. The comparison showed that the shape-DNA technique is capable of cleaning age-related degradations and smudge marks in real-world documents. Also, as shown in Figure 1, on the controlled set of handwritten documents the shape-DNA technique was able to recover most of the missing pieces of the broken characters. In coming months, we will train shape databases on handwritten data collected under the MADCAT program to better model the shape variability in handwritten text.

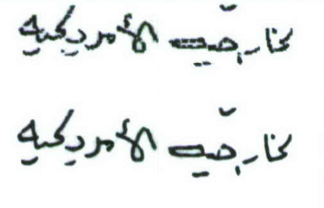

**Figure 1: Restoration of Handwritten text.**

## 1.2 Page Segmentation [BBN, Argon, Lehigh, Polar Rain, UMD, SUNY]

The objective of the page segmentation task is to decompose an image into "perceptually consistent" zones, where perceptual consistency is described in terms of local texture similarities, with the goal of producing zones that contain a single type of content (machine-printed text, handwritten text, graphics, etc.).

**Pixel Level Zoning [Lehigh]**: During this period, we started investigating pixel-based approaches for page segmentation that, while computationally demanding, offer the potential to be general and robust. Our techniques employ a range of classifier technologies, including brute-force k-Nearest Neighbors (kNN), fast approximate kNN using hashed k-d trees, classification and regression trees, and locality-sensitive hashing. Initial experiments suggest that per-pixel accuracies are modest, in the range of 60% to 70%.

**Handwritten Line Detection [UMD]**: In this reporting period, we tested our existing level-set based handwritten line detection on the sample images from the Harmony corpus. We found that, for the documents against which we tested, overlapping text lines, interaction with graphics and noise and large variations in text size all cause significant problems. In the next quarter, we will focus on addressing these problems.

**Texture-based Text Detection [Argon]**: Our approach for text detection in camera-captured images is based on measuring shape similarity of texture. This quarter, we developed an initial version of the text detection module. For training, we have implemented a graphical user interface to allow for manual construction of ground-truth data. The tool ingests an image, partitions it into super pixels (collection of similar pixels), and classifies each super pixel. These classifications can then be manually corrected, and the corrections fed back in as new examples to train the support vector machine (SVM) classifier. In the coming months, we will run experiments to measure accuracy of the text detection module.

## 1.3 Text Recognition [BBN, Argon, Columbia, SUNY]

**Baseline Experiments with HMMs [BBN]**: This quarter, we developed the hardware and the software infrastructure for performing text recognition experiments. Since no MADCAT data has been distributed to date, we validated our experimental setup on the existing corpora we have in-house: the 1995 DARPA Arabic machine-print (DAMP) corpus consisting of 345 images from newspapers, books, magazines, etc. and the IFN/ENIT corpus consisting of 26459 images of handwritten (HW) Tunisian addresses. For experiments on the DAMP corpus, we used 14-state context-independent HMMs for modeling the contextual form of Arabic characters and a word or character n-gram language model (LM). As shown in Table 1, the word error rate (WER) on the DAMP test set was 18.3% and 14.2% with a word trigram and character trigram LM, respectively. The WER for the word LM is higher because of the high (16.4%) out-of-vocabulary rate (OOV) with a 65K decoding lexicon. Note that the OOV rate and the WER are overestimated because we do not detach punctuation characters from the word. In the coming months, we will revisit the WER scoring and OOV computation.

For our baseline experiments on the IFN/ENIT HW corpus, we used context-dependent HMMs and a "compound word" LM to model address strings as a single token. The string accuracy on the IFN/ENIT test set was 95.7%, which is superior to the best published results on that data set.

**Improved Language Models [BBN]**: We explored the following LM improvements in this reporting period:

1. Higher order n-gram rescoring of N-best lists with n > 3.
2. Parts-of-Arabic-Word (PAW) n-gram LMs to model wider context than character LMs while still preserving the unlimited word vocabulary attribute of character n-grams.
3. Knowledge based rescoring of N-best lists using a Finite State Machine (FSM) that models the syntax of contextual forms of Arabic characters.

| Language Model | %WER |
|---|---|
| Word 3-gram | 18.3 |
| Char 3-gram | 14.2 |
| Char 5-gram | 13.2 |
| **PAW 3-gram** | **12.1** |

**Table 1: WER on DAMP corpus**

As shown in Table 1, both the PAW trigram and the character 5-gram LM result in significantly lower WER than the WER obtained from using character or word trigrams. Our initial FSM-based rescoring of N-best lists did not yield any improvement on the DAMP or IFN/ENIT corpus. Next quarter, we will revisit the FSM-based rescoring on the recently distributed Applied Media Analytics (AMA) HW collection and other available MADCAT data.

**Novel Feature Extraction [SUNY, BBN]**: This quarter, we developed an initial version of a standalone library for extracting Gradient-Shape-Concavity (GSC) features for integration into the BBN Byblos system. GSC features combine three different attributes – the gradient representing the local orientation of strokes; structural features that extend the gradient to longer distances and provide information about stroke trajectories; and concavity that captures stroke relationships at long distances.

In this period, we also developed a standalone library for chain-code feature extraction. The chain code feature extraction module traces the chain-code and identifies sharp turning points, end points, crossing points, and connection points in the glyphs under consideration. In the coming months, we will perform recognition experiments with GSC and chain-code features augmented to the standard percentile-based feature stream of the BBN Byblos system.

Moments are another set of features we explored within the BBN Byblos system. Specifically, we experimented with Hu moments, which belong to the family of moments that are scale, rotation, and translation invariant. In our preliminary experiments, we augmented the 7 Hu moments to the existing percentile feature stream and then transformed the resulting 40-dimension vector to 15 dimensions using LDA. Next, we trained glyph HMMs on the DAMP training data. Recognition experiments did not yield an improvement in the WER over the baseline. These experiments were performed with almost no changes to the configuration for feature extraction. Specifically, the frame width chosen may not be optimal for Hu moments feature stream. Therefore, in the next quarter, we will optimize the configuration for experiments with Hu moments.

**PAW based Glyph HMMs [BBN]**: Although character HMMs have been shown to work well with features that capture local variations, we hypothesize that modeling units such as PAWs that are wider than characters should be better suited for capturing structural variations. To investigate that hypothesis, we setup initial experiments for modeling PAWs explicitly. In addition to the 162 characters trained in the baseline experiment on the DAMP corpus, we modeled the 40 most frequent PAWs. The PAWs HMMs used number of states proportional to the number of constituent characters. That is, a PAW consisting of 2 characters was modeled by 28 state HMMs; a PAW consisting of 3 characters was modeled by 42 state HMMs, and so on. The initial results on the DAMP test set resulted in a WER of 12.7%, which is 0.6% absolute worse than the WER obtained using character HMMs. We believe this degradation in WER can be due to two reasons. First, we used the standard features that are designed for capturing local variations rather than longer-span structural characteristics. Second, these experiments were performed on machine-print data which exhibits almost no variations in the rendering of the PAWs when compared to handwritten text. In the coming months, we will revisit PAW HMMs for modeling structural feature streams and applying them to handwritten text.

**Discriminative Training for Glyph HMMs [BBN]**: This reporting period, we experimented with discriminative training for improving our glyph HMMs. Specifically, we used character lattices generated using the baseline models trained with maximum likelihood criterion to perform maximum mutual information (MMI) training. Next, we decoded the DAMP test set with MMI models and a PAW LM. As shown in Table 2, the models trained using MMI outperform the ML models by 1.3% absolute in WER.

| Model Type | %WER |
|---|---|
| ML | 12.1 |
| MMI | 10.8 |

**Table 2: Results with MMI on DAMP test set.**

**Hypotheses Combination using NIST ROVER [BBN]**: Given that the PLATO team is investigating complementary approaches for text recognition, we seek to improve the recognition accuracy by combining the outputs of these complementary techniques. In this period, we setup initial experiments for combining different system hypotheses using the NIST ROVER algorithm. We only combined results from different configurations of the BBN Byblos system, but the setup can easily scale to combine outputs from a large number of systems. Combining 5 different system outputs resulted in a WER of 11.4% on the DAMP test set, a 0.7% absolute improvement over the best configuration that had a WER of 12.1%.

**Probabilistic Bipartite Graph Matching [Argon]**: Building on the bipartite graph matching (BGM) approach applied to OCR in the document understanding seedling, our focus in this quarter was to extend the basic BGM approach to be machine trainable. Through the use of kernel methods and Relevance Vector Machines (RVM), we developed a mathematically rigorous approach to probabilistic graph comparison. We have also started implementing a hyperkernel optimization routine using Brent's method in multiple dimensions to allow automated calculation of hyperkernel parameters.

## 1.4 Metadata Extraction and Layout Interpretation [BBN, BAE, Lehigh, SUNY, UMD]

In the project kickoff meeting held at BBN on 13 November 2007, we were directed by the DARPA Program Manager to perform only limited work in metadata extraction for the initial phase of the MADCAT project. Specifically, our work in metadata extraction is now focused on page segmentation (or zone labeling), text detection, and line finding. During this period, UMD performed experiments to establish a baseline for zone labeling techniques. UMD reported a classification accuracy of 97% on the University of Washington corpus, which has well formed zones.

## 1.5 System Integration and Technology Transition [BBN]

In this period, we held teleconferences and in-person meetings with multiple members of the PLATO team to discuss design of interfaces for different modules of the envisioned PLATO system.